NUQA: Estimating cancer spatial and temporal heterogeneity and evolution through alignment-free methods

Original Article - Methods

CFP-Medicine

NUQA: Estimating cancer spatial and temporal heterogeneity and evolution through alignment-free methods

AC Roddy[1], A Jurek[2], J Souza[1], A Gilmore[1], PG O'Reilly[1], A Stupnikov[1, 3], D Gonzalez de Castro[1], KM Prise[1], M Salto-Tellez[1], DG McArt[1]**

1. Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, UK

2. School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast,

Belfast, UK

3. Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, USA, MD 21287

****Corresponding author:**

Dr. Darragh G. McArt
Bioinformatics Group, Health Sciences Building,
Centre for Cancer Research and Cell Biology,
Queen's University Belfast,
97 Lisburn Road, Belfast, BT9 7AE, UK.
Ph: 0044 (0) 28 9097 2629
Email: d.mcart@qub.ac.uk

Abstract

Background: Longitudinal next generation sequencing of cancer patient samples has enhanced our understanding of the evolution and progression of various cancers. As a result, and due to our increasing knowledge of heterogeneity, such sampling is becoming increasingly common in research and clinical trial sample collections. Traditionally, the evolutionary analysis of these cohorts involves the use of an aligner followed by subsequent stringent downstream analyses. However, this can lead to large levels of information loss due to the vast mutational landscape that characterises tumour samples.

Design: Here, we propose an alignment-free approach for sequence comparison - a well-established approach in a range of biological applications including typical phylogenetic classification. Such methods could be used to compare information collated in raw sequence files to allow an unsupervised assessment of the evolutionary trajectory of patient genomic profiles.

Results: In order to highlight this utility in cancer research we have applied our alignment-free approach using a previously established metric, Jensen-Shannon divergence, and a metric novel to this area, Hellinger distance, to two longitudinal cancer patient cohorts in glioma and clear cell renal cell carcinoma using our software, NUQA.

Conclusion: We hypothesise that this approach has the potential to reveal novel information about the heterogeneity and evolutionary trajectory of spatiotemporal tumour samples, potentially revealing early events in tumorigenesis and the origins of metastases and recurrences.

Introduction

Investigating evolution and heterogeneity of a neoplasm can give insight to the nature and origins of therapeutic resistance as well as assist in predicting response to treatment (Greaves and Maley 2012; Turajlic et al. 2018) As a result, and due to the decreasing costs of next-generation sequencing (NGS), there has been a recent increase in longitudinal profiling of patient samples throughout their care leading to a number of high-quality studies (Gerlinger et al. 2014; Johnson et al. 2014; Mazor et al. 2015; Turajlic et al. 2018). However, there are limitations introduced by bulk sequencing of a tumour and a lack of bioinformatic tools to handle these analyses. Phylogenetic reconstruction is commonly used to study evolution in biology, and so, it would be intuitive to apply this to study clonal evolution in cancer (Nowell 1976). However, current studies build phylogenies based on knowledge from only one type of somatic mutation, such as single nucleotide variants (SNVs) and copy number alteration (Gerlinger et al. 2014; Martínez et al. 2015). These methods also require an alignment step to highlight somatic mutations occurring in each sample introducing information loss and bias due to intrinsic issues previously highlighted (Kidd et al. 2010; Rosenfeld et al. 2012; Paten et al. 2017). Similarly, a number of methods have been highlighted previously to measure intratumoural heterogeneity (ITH) including the use of ecology measures of diversity in Barrett's oesophagus, the MEDICC algorithm, PyClone and EXPANDS (Martínez et al. 2015; Schwarz et al. 2015; Andor et al. 2016). However, similar limitations apply here as only one type of somatic alteration is incorporated, such as allele frequency, also requiring the use of an aligner. Additionally, ecological measures, such as identifying the number of clones, can be found relatively easily in '2-dimensional' tumours such as Barrett's oesophagus but this would be difficult to replicate in 3-dimensional tumours.

Alignment-free sequence comparison, defined as any approach calculating similarity/dissimilarity between sequences which does not use or produce alignment, can be used as an alternative approach to address these issues and create holistic patient profiles for assessing evolutionary trajectories and spatiotemporal heterogeneity. It is more sensitive in the context of sequence divergences and robust against genome rearrangement compared to alignment approaches(Vinga 2014; Bernard et al. 2017). These methods can broadly be split into 2 groups: word-based methods and information-theory based methods. Here, we will focus on word-based methods which have recently been shown to have greater accuracy compared to information theory based methods in protein sequence comparison (Zielezinski et al. 2017). The natural efficiency and accuracy of this algorithm has led to its use in many areas including assessing phylogenetic relationships between bacterial and viral genomes, promoter recognition and protein sequence comparison expanding to an extensive list of tools currently available for various applications(Sims et al. 2009; Chattopadhyay et al. 2015; Fan et al. 2015;

Xu et al. 2016), which has been reviewed previously (Zielezinski et al. 2017). However, very few tools can scale to handle the quantity of data as required by longitudinal cancer research cohorts.

Here, we present NUQA (NGS tool for Unsupervised analysis of fastQ using Alignment-free), a framework that utilizes a highly efficient *k*-mer counter, jellyfish, alongside software built in C++ to quickly and efficiently produce alignment-free 'phylogenetic' trees for longitudinal cancer patient cohorts on a standard workstation. In order to ensure this approach is robustly applicable to cancer research cohorts we have assessed a well-known metric, Jensen-Shannon divergence (JSD), which has previously been applied in an alignment-free context (Sims et al. 2009), as well as a novel metric in this space, Hellinger distance (HD).

New Approaches

NUQA was developed using bespoke scripts written in bash and C++ along with pre-built software jellyfish (Marçais and Kingsford 2011) and phylip(Felsenstein 2004). This algorithm consists of 5 steps: k-mer counting using jellyfish; sorting the resulting count vectors for easier processing and normalising to values between one and zero for comparison; merging the count vectors into a single data matrix using a C++ script; calculating the distances between these vectors using a bespoke C++ script and finally, building a newick tree using phylip. These steps are combined in a single wrapper script written in bash (Figure S1). We have tested both JSD and HD for applicability in the comparison of WES samples in longitudinal cancer patient cohorts. Given two probability vectors, *P* and *Q,* JSD is defined as:

$$JS(P,Q) = \frac{1}{2}KL(P,M) + \frac{1}{2}KL(Q,M)$$

where $M = \frac{1}{2}(P+Q)$ and KL is Kullback-Leibler divergence:

$$KL(P,M) = \sum_{i=1}^{k} p_i log_2 \frac{p_i}{m_i}$$

HD is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$$

Detailed methods are available in the supplementary material (Section 1) and the software implementation can be found on GitHub (https://github.com/ACRoddy/NUQA)

.

Results

Identifying optimal parameters

Multiple distance metrics have been highlighted for their utility in alignment-free sequence comparison in various studies and reviews (Höhl et al. 2007; Dai et al. 2008; Vinga 2014; Zielezinski et al. 2017), From these we selected the most applicable to our cohort (discussed in the Supplementary Material; Note 1.2). We decided to focus on JSD, a previously studied metric in alignment-free methods, and HD which is novel to this domain. We applied each of these metrics to 6 patients, 3 clear cell renal cell carcinoma (ccRCC) patients and 3 glioma patients, using a 21-mer length in order to assess their applicability to cancer patient cohorts (Figure 1A-D and S2). We compared the trees using both Branch Score distance (BSD) and Symmetric distance (SD) (Figure 1C and D). BSD suggests that HD produces similar results to JSD with distances <0.3 for 5/6 patients, while SD highlighted that JSD and HD produce the same tree topologies (SD=0) for all patients except P17 which obtained a SD of 2 due to a change in location of sample 'Recurrence A'. We conclude that JSD and HD both produce consistent results in this context suggesting that HD may perform well in other alignment-free applications.

With the aim of identifying an appropriate *k*-mer length which should be used when applying alignment-free methods to longitudinal cancer patient cohorts, we assessed the effect of varying *k*-mer length (13, 15, 17, 19, 21, 23, 25 and 31bp) for patient RMH004 (Figure 1E-G) and additionally for patients P90, P17, EV001 and EV002 (Figures S3, S4, S5 and S6, respectively) using JSD. An ideal *k*-mer length would have the sensitivity of representing only one mutation while also ensuring it does not occur frequently or represent multiple regions (Supplementary Note 1.2).

Again, we compared trees using BSD and SD. Results were visualised using heatmaps (Figure 1F and G) and a line graph depicting the effects of sequential increases in *k*-mer length on BSD (Figure 1H). Results indicate an optimal range of 17 to 25 for these patients supporting previous findings that 21 is an optimal *k*-mer length for large genomes (Sims et al. 2009; Fan et al. 2015).

Application to cancer patient cohorts

To first validate the use of this method on longitudinal, spatial and temporal cohorts, we created simulated datasets, A and B, to represent cancer patient profiles through introducing controlled mutational events (Figure 2A and 2B, respectively). The aim was to anticipate a predefined branching pattern and assess the ability of NUQA to correctly assign a branching pattern. A ‘normal’ file (N) was produced initially before being mutated to form a ‘cancerous’ file (C). This cancerous sample was then mutated 3 separate times to represent heterogeneity (files C1a, C2a and C3a) and finally each of these three files were mutated two successive times (files b and c) to represent the evolution of these 3 subclones. Dataset A was simulated to represent

SNVs and indels within WES data while dataset B represents SNVs, indels and structural variants within WGS data. As expected, these 3 subclones form 3 distinct branches with file 'c' being the most distal sample and file 'a' being the least.

Furthermore, we identified two well-studied, high quality longitudinal cancer research cohorts to test the utility of our software in glioma (Johnson et al. 2014; van Thuijl et al. 2015) and ccRCC (Gerlinger et al. 2014). We have identified one patient from each cohort for whom the original authors have produced phylogenetic trees drawn from the information obtained using a variant caller to highlight SNVs and small indels. Patient P90 from the glioma cohort had longitudinal samples, whole-exome sequenced, including circulating blood samples (Normal), six samples from the initial tumour (Initial A-F) classified as a grade II glioma and 2 samples from a recurrence tumour (Recur A and B) also classified as a grade II glioma. We applied our own algorithm to this patient and produced phylogenetic trees and MDS plots based on our output (Figure 2B and C). A least-squares minimum-evolution (LSME) tree was produced from somatic SNVs and indels for patient P90 by Mazor *et al.*, for which, detailed methods can be found in the original paper (Mazor et al. 2015) (Figure 2A). We use these as a basis for comparison, aware that bias will have been introduced as only reads which uniquely aligned to the reference genome have been considered and the variant callers used could only identify SNVs and small indels but not larger aberrations. This tree contains a relatively long trunk region before tumour samples diverge indicating linear evolution. Furthermore, three key clusters of samples are formed, the first containing Initial C, D and F, the second containing Initial A, B and E and the final cluster containing the two recurrent samples. Similarly, the tree produced using NUQA is highly consistent, also indicating that initial samples C, D and F occur early in evolution, clustering closely with the Normal sample while initial samples A, B and E branch distally suggesting that these are later events in evolution. In addition, recurrence samples A and B branch early, clustering closely with initial samples C, D and F. Moreover, both trees seem to suggest high levels of ITH within the initial tumour and that there is little ITH within the recurrent tumour.

For ccRCC patient RMH004 we have WES data for germline DNA in the blood (GL), 5 samples from the initial ccRCC tumour (R2-4, R8 and R10) and 1 sample from a thrombus in a renal vein (VT). Again, we produced phylogenetic trees and MDS plots based on our output from NUQA for this patient (Figure 2E and F). Maximum parsimony trees were created based on SNVs and small indels found to be present within the tumour samples as described in the original paper (Gerlinger et al. 2014) (Figure 2D). The original maximum parsimony tree suggests that R3, VT and R10 occur early in evolution while R8, R4 and R2 occur much later and are more highly mutated. The original authors highlighted that two distinct mutations occurred in *PBRM1* indicating parallel evolution of two subclones within the tumour. The phylogenetic tree produced using NUQA also suggests that sample R10 occurs early in evolution and that R2 and R4 are more genetically

divergent, occurring much later in evolution (Figure 2E). However, samples R3, VT and R8 show variations in branching suggesting that more complex mutational events may be present in these samples. Both trees also appear to show high levels of ITH which can also be seen in the MDS plot for these samples (Figure 2F).

Further analysis of patients P17 and EV001 also indicate similar groupings to what can be seen using alignment-based methods, however, again there are key differences in branching within these patients (Figure S6;Note 3.1 and S7;Note 3.2, respectively). Additional analyses can be performed based on these results, for example, by using the branching pattern produced through NUQA to inform groups as a basis for further analysis. An example using FastGT (Pajuste et al. 2017) to identify SNP calls differentiating groups found in patient P90 can be found in the supplementary material (Note 3.3).

Benchmarking alternative alignment-free packages

Reviewing the literature on current alignment-free phylogenetic software identified two capable of processing multiple large fastq files for sequence comparison: AAF and kWIP (Fan et al. 2015; Murray et al. 2017) both of which are designed to classify organisms at species level requiring a sensitivity to much larger genetic distances. All packages were tested using patient P90 using a *k*-mer length of 21 and allowing 64 GB RAM. AAF produced the best time of 1 hour, 57 minutes while NUQA ran in 2 hours, 25 minutes and kWIP ran in 5 hours, 48 minutes. In order to assess the applicability of these to cancer research data we tested NUQA, AAF and kWIP on our simulated dataset (Figure 3A, 3B and 3C, respectively). It is promising to see that all softwares produce the branching pattern we expect to see. However, when applied to patient P90 (Figure 3D, E and F, respectively) we see a variation in tree topology, but more importantly, AAF and kWIP produce a very small trunk (orange) compared to branch lengths indicating that they are less sensitive to the changes occurring between single-patient samples.

Discussion

Alignment-free sequence comparison is capable of building evolutionary relationships between samples without the use of an aligner. This approach allows for the inclusion of all information regardless of whether it would align to a reference genome preventing bias to pipeline specific information and allowing the inclusion of larger insertions and deletions or chromosomal rearrangements which would be difficult to align. It is also a highly efficient approach yielding grossly improved times over traditional methods using an aligner (Zielezinski et al. 2017).

Here, we have shown potential utility for this approach to be applied to longitudinal cancer patient cohorts as an unsupervised approach for comparing sequencing files. In order to do this we have tested a range of suitable distance metrics for their applicability to this type of data, highlighting

JSD as an appropriate measure to assess pairwise distances between feature frequency profiles as previously described (Sims et al. 2009). But also HD, a previously untested metric in alignment-free sequence comparison which we have shown produces equally consistent results. Varying *k*-mer length revealed that a *k*-mer greater than 17 should be sufficient for this analysis, however, we decided to continue further analysis with a *k*-mer length of 21 to reduce the effects of homoplasy. We validated the use of NUQA on longitudinal, spatial and temporal cohorts using 2 simulated datasets A and B, representing SNVs and indels in small scale data and SNVs, indels and structural variants in large scale, WGS data, respectively. Furthermore, we assessed the utility of applying an alignment-free framework in cancer research by applying this method to one patient each from two high-quality longitudinal cohorts in ccRCC (Gerlinger et al. 2014) and glioma (Johnson et al. 2014; Mazor et al. 2015). In both cases, clear similarities could be seen when comparing the results of alignment-free analysis to the trees produced using alignment-based approaches, deduced from changes in SNVs and small indels, however, clear and possibly fundamental differences could be seen. This may be a result of unassessed gene fusion events, larger indels or chromosomal rearrangements which are also contributing to the tumours mutational landscape and therefore affecting the evolutionary pathway of these cancer patients. Finally, we benchmarked our software, NUQA, against other large-scale alignment-free softwares designed for assessing a much greater genetic divergence between samples: kWIP and AAF. We found that AAF yielded a marginal improved speed over our current approach, however, neither software was designed to assess the relatively small genetic distances which would be seen in a cancer patient cohort.

Our tool, in combination with alignment-free genotyping tools, such as FastGT (Pajuste et al. 2017), has the potential to add extra layers to the evolutionary analyses of cancer types providing insights which may otherwise be passed over. Further analysis of the feature frequency profiles built in our extendable alignment-free framework could highlight patterns and abnormalities contributing to the branching pattern obtained for each cancer patient helping to tease out contributing factors in cancer evolution. We would expect that given current precision medicine paradigms and reductions in sequencing costs this approach may be adopted clinically to highlight a cancer trajectory and consequential strategies for the patient.

In conclusion, we have introduced NUQA, a novel and efficient software application for performing alignment-free sequencing comparison, with the aim of highlighting the utility of these methods for the unsupervised phylogenetic assessment of longitudinal patient cohorts in cancer research. We hypothesise that this presents an opportunity to provide a landscape view to identify early and late events in evolution as well as give an indication of the origins of metastatic and recurrent tumours in quick turnaround time and can be used in combination with the more targeted and previously adopted approaches.

Materials and Methods

This framework was applied to 2 previous published datasets: A glioma cohort containing spatial and temporal exome-seq data for patients P17, P49 and P90 (Johnson et al. 2014; Mazor et al. 2015), A ccRCC cohort containing spatial and temporal exome-seq data for patients EV001, EV002 and RMH004 (Gerlinger et al. 2014).

Both the glioma and the ccRCC cohort were pre-processed using the same steps prior to applying our algorithm: SAMTOOLs (Li and Durbin 2009) was used to revert files for patient's P17, EV001, EV002 and RMH004 from bam to fastq files to allow us to work with the raw reads obtained from sequencing. Following this, FastQC (Andrews 2010) was used to ensure the files were a good quality for alignment-free processing and for setting levels for trimming, if required reads were trimmed using Trimmomatic (Bolger et al. 2014). Finally, resulting trees were visualised using the online software tool, iTOL (https://itol.embl.de/). MDS plots were created using the *cmdscale()* function and *ggplot2* (Wickham 2009) package within the R statistical environment (R Core Team 2017).

To assess changes in tree topology and branch lengths between trees produced using alignment-free methods for the same patient we used Branch-Score distance, a measure accounting for both branch length and tree topology, and Symmetric distance, a measure accounting for only tree topology. Both of these are available through the Phylip package.

Further description of the generation of simulated data and discussion on the choice of distance metric and evaluation of k-mer length are available in supplementary notes 1.2-1.6

Acknowledgements

This work was supported by funding from the Brainwaves Northern Ireland Charity (Registered Charity Number: NIC103464). AC Roddy is supported by a Cancer Research UK studentship (C11512/A20877). The authors would like to thank Joseph Costello and team for assistance in data transfer and clarity of files. The authors would also like to thank Chris O'Neill for fruitful discussions around application.

References

Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, Ji HP, Maley CC. 2016. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat Med. 22(1):105–113. doi:10.1038/nm.3984.Pan-cancer.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bernard G, Chan CX, Chan Y, Chua X-Y, Cong Y, Hogan JM, Maetschke SR, Ragan MA. 2017. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Brief Bioinform.(March):1–10. doi:10.1093/bib/bbx067.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

Chattopadhyay AK, Nasiev D, Flower DR. 2015. Sequence analysis A statistical physics perspective on alignment- independent protein sequence comparison. Bioinformatics. 31(March):2469–2474. doi:10.1093/bioinformatics/btv167.

Dai Q, Yang Y, Wang T. 2008. Markov model plus k-word distributions: A synergy that produces novel statistical measures for sequence comparison. Bioinformatics. 24(20):2296–2302. doi:10.1093/bioinformatics/btn436.

Fan H, Ives AR, Surget-Groba Y, Cannon CH. 2015. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics. 16(1):1–18. doi:10.1186/s12864-015-1647-5.

Felsenstein J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. http://www.evolution.gs.washington.edu/phylip.html. [accessed 2018 Apr 9]. http://ci.nii.ac.jp/naid/10027221536/en/.

Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, Mcgranahan N, Matthews N, Santos CR, et al. 2014. Articles Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Publ Gr. 46(3):225–233. doi:10.1038/ng.2891.

Greaves M, Maley CC. 2012. Clonal evolution in cancer. Nature. 481(7381):306–313. doi:10.1038/nature10762.

Höhl M, Rigoutsos I, Ragan MA. 2007. Pattern-based phylogenetic distance estimation and tree reconstruction. Evol Bioinform Online. 2(2003):359–75. doi:10.1080/10635150701294741.

Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K, et al. 2014. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science (80-). 343(6167):189–193. doi:10.1126/science.1239947.

Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G, et al. 2010. Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods. 7(5):365–371. doi:10.1038/nmeth.1451.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754–60. doi:10.1093/bioinformatics/btp324.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27(6):764–770. doi:10.1093/bioinformatics/btr011.

Martínez E, Yoshihara K, Kim H, Mills GM, Treviño V, Verhaak RGW. 2015. Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects. Oncogene. 34(21):2732–2740. doi:10.1038/onc.2014.216.

Mazor T, Pankov A, Johnson BE, Hong C, Hamilton EG, Bell RJA, Smirnov I V., Reis GF, Phillips JJ, Barnes MJ, et al. 2015. DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. Cancer Cell. 28(3):307–317. doi:10.1016/j.ccell.2015.07.012.

Murray KD, Webers C, Ong CS, Borevitz J, Warthmann N. 2017. kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. PLoS Comput Biol. 13(9):1–15. doi:10.1371/journal.pcbi.1005727.

Nowell PC. 1976. The clonal evolution of tumor cell populations. Science (80-). 194(4260):23–28. doi:10.1126/science.959840.

Pajuste FD, Kaplinski L, Möls M, Puurand T, Lepamets M, Remm M. 2017. FastGT: An alignment-free method for calling common SNVs directly from raw sequencing reads. Sci Rep. 7(1):1–10. doi:10.1038/s41598-017-02487-5.

Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. Genome Res. 27(5):665–676. doi:10.1101/gr.214155.116.

Rosenfeld JA, Mason CE, Smith TM. 2012. Limitations of the human reference genome for personalized genomics. PLoS One. 7(7):1–9. doi:10.1371/journal.pone.0040294.

Schwarz RF, Ng CKY, Cooke SL, Newman S, Temple J, Piskorz AM, Gale D, Sayal K, Murtaza M, Baldwin J, et al. 2015. Spatial and Temporal Heterogeneity in High- Grade Serous Ovarian Cancer : A Phylogenetic Analysis. PLoS Med. 12(2):1–20. doi:10.1371/journal.pmed.1001789.

Sims GE, Jun S-R, Wu GA, Kim S-H. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci. 106(8):2677–2682. doi:10.1073/pnas.0813249106.

Team RC. 2017. R: A Language and Environment for Statistical Computing.

van Thuijl HF, Mazor T, Johnson BE, Fouse SD, Aihara K, Hong C, Malmström A, Hallbeck M, Heimans JJ, Kloezeman JJ, et al. 2015. Evolution of DNA repair defects during malignant progression of low-grade gliomas after temozolomide treatment. Acta Neuropathol. 129(4):597–607. doi:10.1007/s00401-015-1403-6.

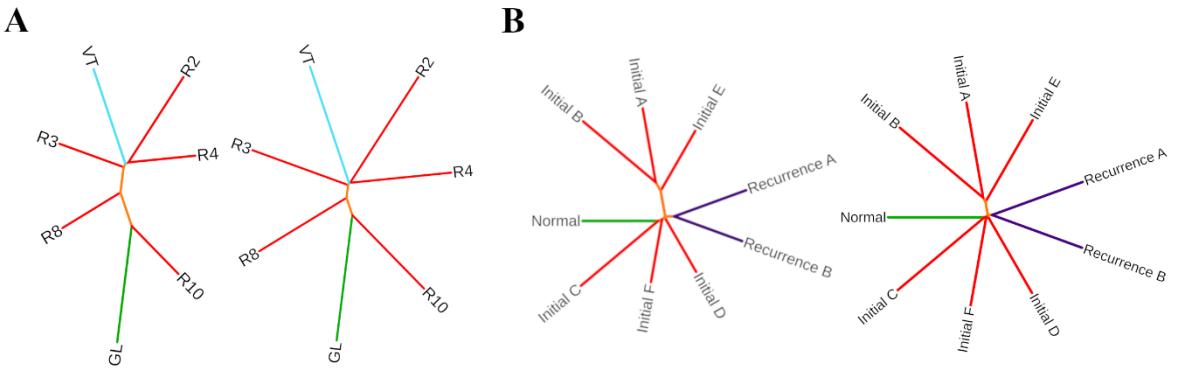
Turajlic S, Xu H, Litchfield K, Rowan A, Chambers T, Lopez JI, Nicol D, O'Brien T, Larkin J, Horswell S, et al. 2018. Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. Cell. 173(3):581-594.e12. doi:10.1016/j.cell.2018.03.057.

Vinga S. 2014. Information theory applications for biological sequence analysis. Brief Bioinform. 15(3):376–389. doi:10.1093/bib/bbt068.

Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

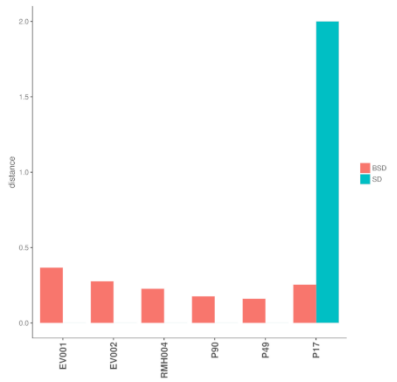
Xu W, Zhang L, Lu Y. 2016. SD-MSAEs : Promoter recognition in human genome based on deep feature extraction. J Biomed Inform. 61:55–62. doi:10.1016/j.jbi.2016.03.018.

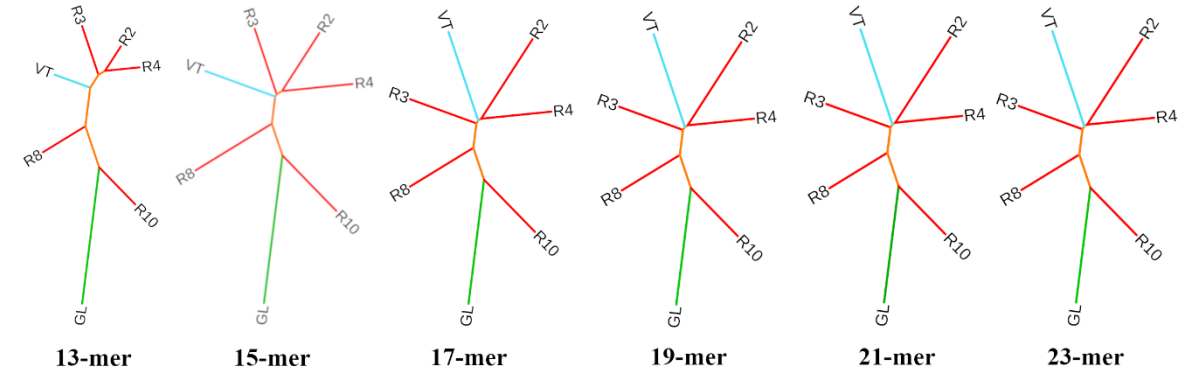
Zielezinski A, Vinga S, Almeida J, Karlowski WM. 2017. Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biol. 18(1):1–17. doi:10.1186/s13059-017-1319-7.



C

Patient	BSD	Symmetric
EV001	0.36704	0
EV002	0.27558	0
RMH004	0.22733	0
P90	0.17661	0
P49	0.16069	0
P17	0.25407	2





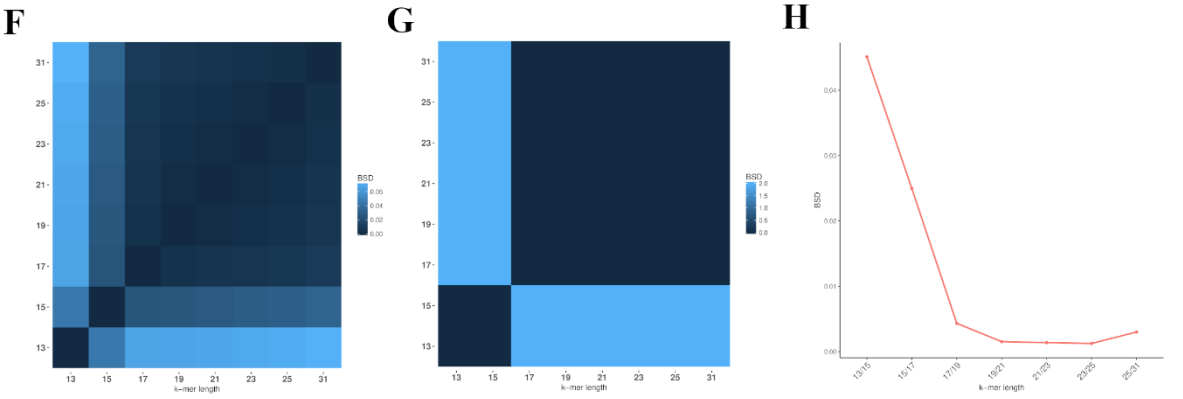


Figure 1: Identifying optimal parameters for use with alignment-free. Application of Jensen-Shannon Divergence (JSD) and Hellinger Distance (HD) to (**A**) clear cell renal cell carcinoma (ccRCC) patient RMH004 with a germ-line sample (GL), multiple samples from the ccRCC tumour (R2-4, R8, R10) and a tumour thrombus from the renal vein (VT) and (**B**) glioma patient P90 with a germ-line sample (Normal), multiple samples from the initial grade II glioma (Initial A-F) and 2 samples from a recurrent grade II glioma (Recur 1A and 1B). (**C**) A table summarizing Branch score distance (BSD) and Symmetric distance (SD) values returned when comparing trees for 6 patients for which both JSD and HD have been applied. (**D**) A bar chart summarizing BSD and SD values returned when comparing trees for 6 patients for which both JSD and HD have been applied. (**E**) Tree topologies produced using *k*-mer lengths 13, 15, 17, 19, 21 and 23 in combination with JSD when applying alignment-free methods to patient RMH004. (**F**) A heatmap representing the Branch-Score distance (BSD) between trees produced using varying *k*-mer lengths and HD applied to patient RMH004. (**G**) A heatmap representing the BSD between trees produced using varying *k*-mer lengths and JSD applied to patient RMH004. (**H**) A line graph representing the BSD between trees produced using increasing *k*-mer lengths when applying JSD.

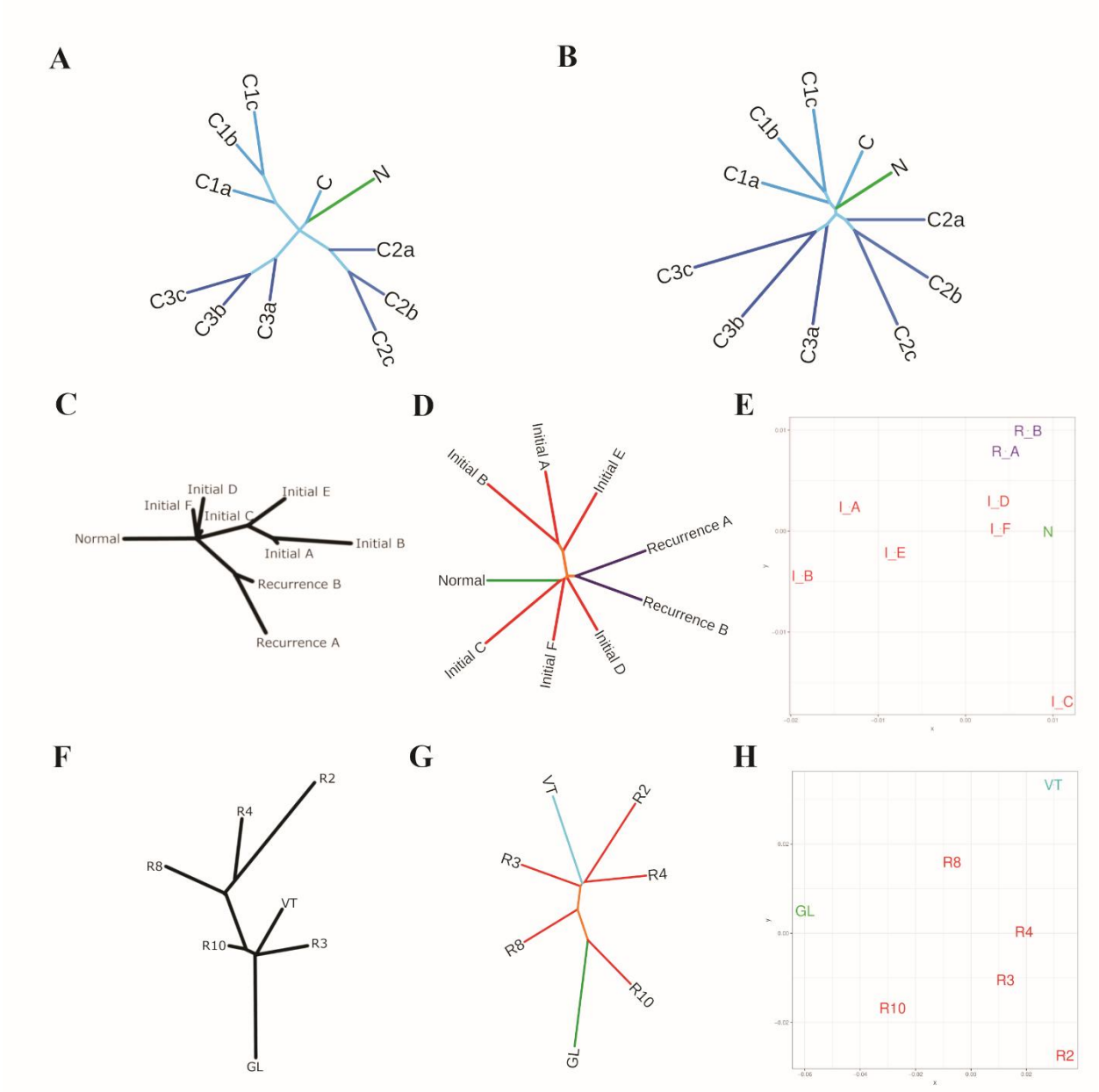


Figure 2: Applying alignment-free sequence comparison methods to glioma patient P90 and ccRCC patient RMH004. (**A**) Simulated dataset 'A' created using software XS and fastx-mutate-tools to represent SNVs and indels in small scale data such as WES (**B**) Simulated dataset 'B' created using software pIRS to represent SNVs and indels and structural variants in WGS (**C**) Least-square minimum-evolution tree produced based on a binary matrix of SNVs present in the samples for P90 adapted from Mazor *et al*. (**D**) An unrooted neighbour-joining tree produced applying our alignment-free software (NUQA), incorporating JSD, to patient P90. (**E**) Multi-dimensional scaling plot representing the distances between samples produced applying NUQA, incorporating JSD, to patient P90. (**F**) A maximum parsimony tree produced based on a binary matrix of SNVs present in the samples for RMH004 adapted from Gerlinger *et al*. (**G**) An unrooted neighbour-joining tree produced applying NUQA, incorporating JSD, to patient RMH004. (**H**) Multi-

dimensional scaling plot representing the distances between samples produced applying NUQA, incorporating JSD, to patient RMH004.

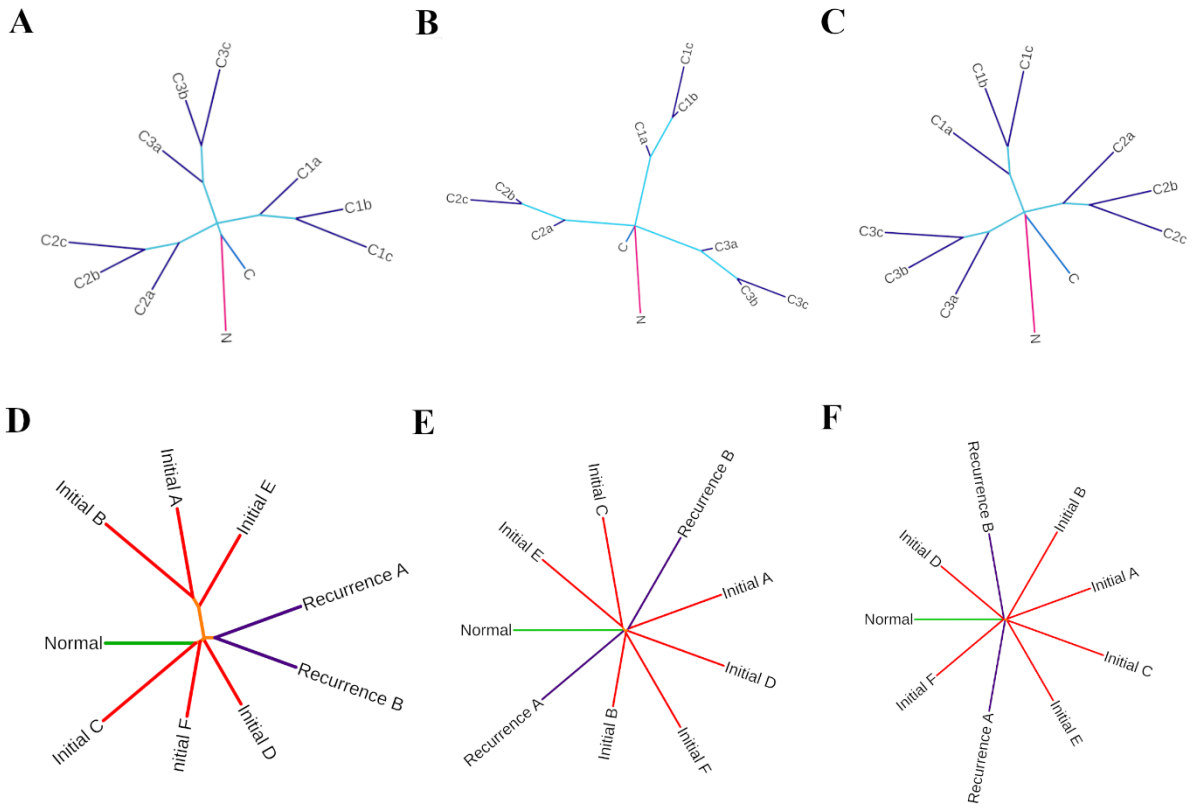


Figure 3: Benchmarking of NUQA against other alignment-free softwares. Unrooted neighbour-joining trees produced when applying NUQA (**A**), AAF (**B**) and kWIP (**C**) to a simulated datset using a *k*-mer length of 17 and allowing 64GB RAM and trees produced when applying NUQA (**D**), AAF (**E**) and kWIP (**F**) to patient P90 using a *k*-mer length of 21 and allowing 64GB RAM.